CLAIMS

1. A method comprising:

obtaining a set of training data having associated summaries;

using the set of training data and associated summaries to generate a key feature generation model;

obtaining another set of training data having associated categories;

mapping, using the key feature generation model, the other set of training data to a set of vectors; and

training a data classifier based on the set of vectors and the associated categories.

2. A method as recited in claim 1, further comprising:

receiving data to be classified;

using the key feature generation model to obtain a vector representing the data to be classified;

inputting the obtained vector to the trained data classifier; and

obtaining, from the trained data classifier, a category in which the data is classified, wherein the category is one of the associated categories.

3. A method as recited in claim 1, wherein the training data comprises a plurality of pieces of training text, wherein the associated summaries include keywords, and wherein at least one summary corresponds to each piece of training text.

4. A method as recited in claim 1, wherein the set of training data and the other set of training data comprise the same data.

5. A method as recited in claim 1, wherein using the set of training data and associated summaries to generate the key feature generation model comprises:

obtaining, from the associated summaries, a key feature listing; and

creating, for each key feature in the key feature listing, a key feature classifier which judges how likely it is for the key feature to occur in one of the associated summaries.

6. A method as recited in claim 5, wherein obtaining the key feature listing comprises including, in the key feature listing, each key feature that is present in at least one of the associated summaries.

7. A method as recited in claim 5, wherein the key feature classifier comprises a naïve Bayesian classifier.

8. A method as recited in claim 5, wherein the key feature classifier comprises a probabilistic model.

9. A method as recited in claim 5, wherein mapping the other set of training data to a set of vectors comprises:

generating a probability vector for each of a plurality of pieces of training data of the other set of training data, wherein each component of the probability

vector for a piece of training data represents the conditional probability of a key feature of the key feature listing given the piece of training data; and

including each generated probability vector as a vector of the set of vectors.

10. A method as recited in claim 5, wherein mapping the other set of training data to a set of vectors comprises:

for each piece of training data, using the created key feature classifiers to generate the elements of a vector of the set of vectors.

11. A method as recited in claim 1, wherein mapping the other set of training data to a set of vectors comprises:

generating a probability vector for each of a plurality of pieces of training data of the other set of training data, wherein each component of the probability vector for a piece of training data represents the conditional probability of a key feature given the piece of training data; and

including each generated probability vector as a vector of the set of vectors.

12. A method as recited in claim 1, wherein the data classifier comprises a support vector machine classifier.

13. A method as recited in claim 1, wherein the data classifier comprises nearest neighbor classifier.

14. A method as recited in claim 1, wherein the data classifier comprises a neural network classifier.

15. A method as recited in claim 1, wherein the data classifier comprises a naïve Bayesian classifier.

16. A method as recited in claim 1, wherein the data classifier comprises a logistic regression classifier.

17. A method as recited in claim 1, wherein the data classifier comprises a rule-based classifier.

18. One or more computer readable media having stored thereon a plurality of instructions that, when executed by one or more processors of a device, causes the one or more processors to:

obtain a set of training text having associated summaries;

use the set of training text and associated summaries to generate a keyword generation model;

obtain another set of training text having associated categories;

map, using the keyword generation model, the other set of training text to a set of vectors; and

train a text classifier based on the set of vectors and the associated categories.

19. One or more computer readable media as recited in claim 18, wherein the instructions that cause the one or more processors to use the set of training text and associated summaries to generate the keyword generation model comprise instructions that cause the one or more processors to:

obtain, from the associated summaries, a keyword listing; and

create, for each keyword in the keyword listing, a keyword classifier which indicates how likely it is for the keyword to occur in one of the associated summaries.

20. One or more computer readable media as recited in claim 19, wherein the instructions that cause the one or more processors to map the other set of training text to a set of vectors comprise instructions that cause the one or more processors to:

generate a probability vector for each of a plurality of pieces of training text of the other set of training data, wherein each component of the probability vector for a piece of training text represents the conditional probability of a keyword of the keyword listing given the piece of training text; and

include each generated probability vector as a vector of the set of vectors.

21. A method of classifying data, the method comprising:

receiving data to be classified;

using a key feature generation model to obtain a vector representing the data, wherein the key feature generation model is based on a set of training data having associated summaries; and

inputting the obtained vector to a trained data classifier, wherein the trained data classifier was previously trained using the set of training data and associated summaries.

22. A method as recited in claim 21, wherein different pieces of training data are used as a basis for the key feature generation model and the trained data classifier.

23. A method as recited in claim 21, wherein the set of training data comprises a plurality of pieces of training text, wherein the associated summaries include keywords, and wherein at least one summary corresponds to each piece of training text.

24. A method as recited in claim 21, wherein the key feature generation model was previously generated by:

obtaining, from the associated summaries, a key feature listing; and

creating, for each key feature in the key feature listing, a key feature classifier which indicates how likely it is for the key feature to occur in one of the associated summaries.

25. One or more computer readable media having stored thereon a plurality of instructions that, when executed by one or more processors of a device, causes the one or more processors to:

train a text classifier using multiple pieces of training text, a plurality of summaries wherein each of the plurality of summaries is associated with one of

the multiple pieces of training text, and a plurality of categories wherein each of the plurality of categories is associated with one of the multiple pieces of training text; and

use the trained text classifier to classify input text without an associated summary.

26. One or more computer readable media as recited in claim 25, wherein the instructions that cause the one or more processors to train the text classifier cause the one or more processors to:

obtain, from the associated summaries, a keyword listing;

create, for each keyword in the keyword listing, a keyword classifier which indicates how likely it is for the keyword to occur in one of the associated summaries; and

use, the created classifiers as a key feature generation model.

27. One or more computer readable media as recited in claim 26, wherein the instructions that cause the one or more processors to obtain the keyword listing comprise instructions that cause the one or more processors to include, in the keyword listing, each keyword that is present in at least one of the associated summaries.

28. One or more computer readable media as recited in claim 26, wherein each of the created keyword classifiers comprises a probabilistic model.

29. One or more computer readable media as recited in claim 26, wherein the instructions that cause the one or more processors to train the text classifier further cause the one or more processors to:

generate a probability vector for each of a plurality of pieces of the multiple pieces of training text, wherein each component of the probability vector for a piece of training text represents the conditional probability of a keyword of the keyword listing given the piece of training text;

include each generated probability vector as a vector of a set of vectors; and

map another plurality of pieces of the multiple pieces of training text to the set of vectors.

30. A system comprising:

a stochastic key feature generation model training module to generate a trained model based on a first training set, wherein the first training set includes training data and associated summaries;

a training data mapping module to generate a plurality of vectors based on the trained model and a second training set, wherein the second training set includes training data and associated categories; and

a classifier training module to construct a trained classifier based on the plurality of vectors and the second training set.

31. A system as recited in claim 30, further comprising:

a stochastic key feature generation model-based vector generation module to generate a vector based on input data and the trained model; and

wherein the trained classifier is to receive the vector and, based on the vector, classify the input data into one or more classes.

32. A system as recited in claim 30, wherein the training data included in the first training set and the training data included in the second training set are the same training data.

33. A system as recited in claim 30, wherein the training data included in the first training set comprises a plurality of pieces of training text, wherein the associated summaries include keywords, and wherein at least one summary corresponds to each piece of training text.

34. A system as recited in claim 30, wherein the stochastic key feature generation model training module is to generate the trained model by:

obtaining, from the associated summaries, a key feature listing; and

creating, for each key feature in the key feature listing, a key feature classifier which judges how likely it is for the key feature to occur in one of the associated summaries.

35. A system as recited in claim 34, wherein the stochastic key feature generation model training module is to obtain the key feature listing by including, in the key feature listing, each key feature that is present in at least one of the associated summaries.

36. A system as recited in claim 34, wherein the key feature classifier comprises a probabilistic model.

37. A system as recited in claim 34, wherein the training data mapping module is to generate the plurality of vectors by:

generating a probability vector for each of a plurality of pieces of training data in the second training set, wherein each component of the probability vector for a piece of training data represents the conditional probability of a key feature of the key feature listing given the piece of training data; and

including each generated probability vector as a vector of the plurality of vectors.

38. A system as recited in claim 34, wherein the training data mapping module is to generate the plurality of vectors by:

using, for each piece of training data, the created key feature classifiers to generate the elements of a vector of the plurality of vectors.

39. A system as recited in claim 30, wherein the training data mapping module is to generate the plurality of vectors by:

generating a probability vector for each of a plurality of pieces of training data included in the second training set, wherein each component of the probability vector for a piece of training data represents the conditional probability of a key feature given the piece of training data; and

including each generated probability vector as a vector of the plurality of vectors.

40. A system comprising:

a stochastic key feature generation model-based vector generation module to generate a vector based on input data and a stochastic key feature generation model, wherein the stochastic key feature generation model was previously generated based on training data and associated summaries; and

a classifier to receive the vector and, based on the vector, classify the input data into one or more classes.

41. A system as recited in claim 40, wherein the training data comprises a plurality of pieces of training text, wherein the associated summaries include keywords, and wherein at least one summary corresponds to each piece of training text.

42. A system as recited in claim 40, wherein the stochastic key feature generation model was previously generated by:

obtaining, from the associated summaries, a key feature listing; and

creating, for each key feature in the key feature listing, a key feature classifier which indicates how likely it is for the key feature to occur in one of the associated summaries.

43. A system comprising:

means for generating a trained model based on a first training set, wherein the first training set includes training data and associated summaries;

means for generating a plurality of vectors based on the trained model and a second training set, wherein the second training set includes training data and associated categories; and

means for constructing a trained classifier based on the plurality of vectors and the second training set.

44. A system as recited in claim 43, further comprising:

means for generating a vector based on input data and the trained model; and

wherein the trained classifier is to receive the vector and, based on the vector, classify the input data into one or more classes.